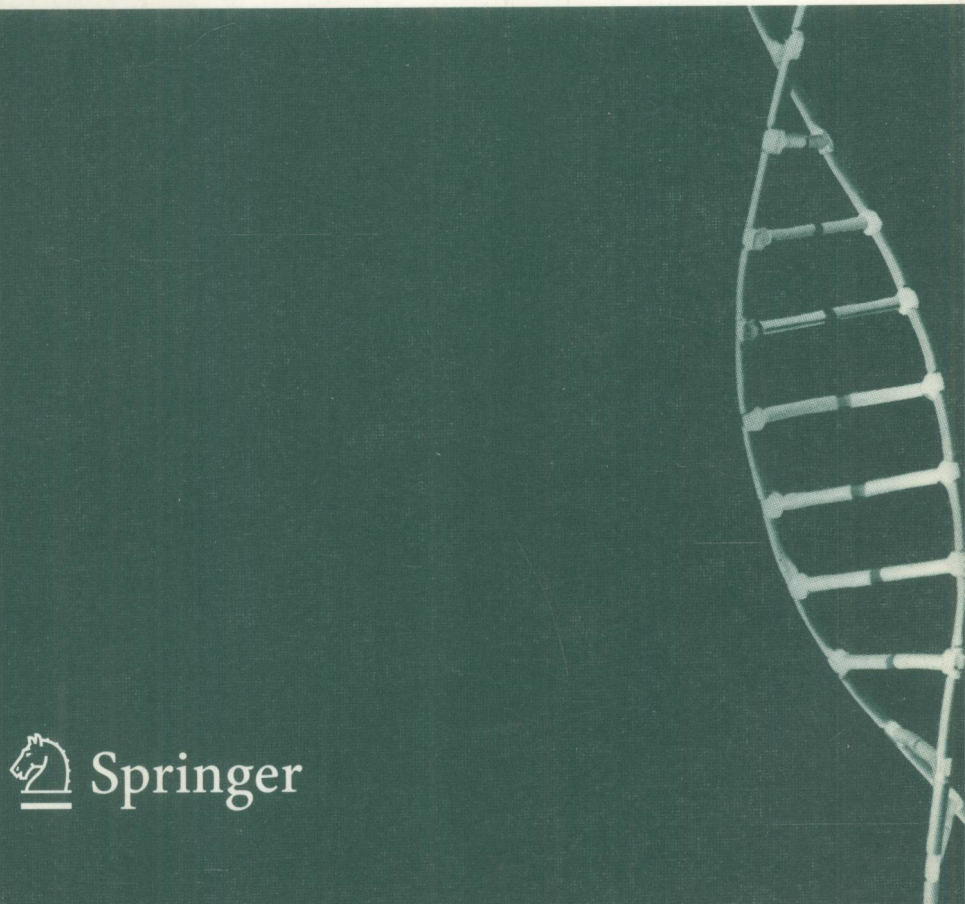
José Luis Oliveira
Víctor Maojo
Fernando Martin-Sanchez
António Sousa Pereira (Eds.)

Biological and Medical Data Analysis

LNBI 3745

6th International Symposium, ISBMDA 2005
Aveiro, Portugal, November 2005
Proceedings



Springer

José Luis Oliveira
Víctor Maojo
Fernando Martin-Sanchez
António Sousa Pereira (Eds.)

Biological and Medical Data Analysis

6th International Symposium, ISBMDA 2005
Aveiro, Portugal, November 10-11, 2005
Proceedings

Springer

Series Editors

Sorin Istrail, Brown University, Providence, RI, USA
Pavel Pevzner, University of California, San Diego, CA, USA
Michael Waterman, University of Southern California, Los Angeles, CA, USA

Volume Editors

José Luis Oliveira
António Sousa Pereira
University of Aveiro
Department of Electronics and Telecommunications (DET/IEETA)
Campus Santiago, 3810 193 Aveiro, Portugal
E-mail: {jlo,asp}@det.ua.pt

Víctor Maojo
Polytechnical University of Madrid
School of Computer Science, Artificial Intelligence Lab
Boadilla del Monte, 28660 Madrid, Spain
E-mail: vmaojo@infomed.dia.fi.upm.es

Fernando Martin-Sanchez
Institute of Health Carlos III
Department of Medical Bioinformatics
Ctra. Majadahonda a Pozuelo, km. 2, 28220 Majadahonda, Madrid, Spain
E-mail: fmartin@isciii.es

Library of Congress Control Number: 2005934196

CR Subject Classification (1998): H.2.8, I.2, H.3, G.3, I.5.1, I.4, J.3, F.1

ISSN	0302-9743
ISBN-10	3-540-29674-3 Springer Berlin Heidelberg New York
ISBN-13	978-3-540-29674-4 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Olgun Computergrafik
Printed on acid-free paper SPIN: 11573067 06/3142 5 4 3 2 1 0

Lecture Notes in Bioinformatics 3745

Edited by S. Istrail, P. Pevzner, and M. Waterman

Subseries of Lecture Notes in Computer Science

Preface

The sequencing of the genomes of humans and other organisms is inspiring the development of new statistical and bioinformatics tools that we hope can modify the current understanding of human diseases and therapies. As our knowledge about the human genome increases so does our belief that to fully grasp the mechanisms of diseases we need to understand their genetic basis and the proteomics behind them and to integrate the knowledge generated in the laboratory in clinical settings. The new genetic and proteomic data has brought forth the possibility of developing new targets and therapies based on these findings, of implementing newly developed preventive measures, and also of discovering new research approaches to old problems.

To fully enhance our understanding of disease processes, to develop more and better therapies to combat and cure diseases, and to develop strategies to prevent them, there is a need for synergy of the disciplines involved, medicine, molecular biology, biochemistry and computer science, leading to more recent fields such as bioinformatics and biomedical informatics.

The 6th International Symposium on Biological and Medical Data Analysis aimed to become a place where researchers involved in these diverse but increasingly complementary areas could meet to present and discuss their scientific results.

The papers in this volume discuss issues from statistical models to architectures and applications to bioinformatics and biomedicine. They cover both practical experience and novel research ideas and concepts.

We would like to express our gratitude to all the authors for their contributions to preparing and revising the papers as well as the Technical Program Committee who helped put together an excellent program for the conference.

November 2005

José Luís Oliveira
Víctor Maojo
Fernando Martín-Sánchez
António Sousa Pereira

Organization

General Chair

José Luís Oliveira, Univ. Aveiro, Portugal

Scientific Committee Coordinators

V. Maojo, Univ. Politecnica de Madrid, Spain
F. Martín-Sánchez, Institute of Health Carlos III, Spain
A. Sousa Pereira, Univ. Aveiro, Portugal

Steering Committee

R. Brause, J.W. Goethe Univ., Germany
D. Polónia, Univ. Aveiro, Portugal
F. Vicente, Institute of Health Carlos III, Spain

Scientific Committee

A. Babic, Univ. Linkoping, Sweden
R. Baud, Univ. Hospital of Geneva, Switzerland
V. Breton, Univ. Clermont-Ferrand, France
J. Carazo, Univ. Autonoma of Madrid, Spain
A. Carvalho, Univ. São Paulo, Brazil
P. Cinquin, Univ. Grenoble, France
W. Dubitzky, Univ. Ulster, UK
M. Dugas, Univ. Munich, Germany
P. Ghazal, Univ. Edinburgh, UK
R. Guthke, Hans-Knoell Institut, Germany
O. Kohlbacher, Univ. Tübingen, Germany
C. Kulikowski, Rutgers Univ., USA
P. Larranaga, Univ. Basque Country, Spain
N. Maglaveras, Univ. Thessaloniki, Greece
L. Ohno-Machado, Harvard Univ., USA
F. Pinciroli, Politecnico di Milano, Italy
D. Pisanelli, ISTC-CNR, Italy
G. Potamias, ICS-FORTH, Greece
M. Santos, Univ. Aveiro, Portugal
F. Sanz, Univ. Pompeu Fabra, Spain
W. Sauerbrei, Univ. Freiburg, Germany
S. Schulz, Univ. Freiburg, Germany

A. Silva, Univ. Aveiro, Portugal
T. Solomonides, Univ. West of England, UK
B. Zupan, Univ. Ljubljana, Slovenia
J. Zvárová, Univ. Charles, Czech Republic

Special Reviewers

G. Moura, Univ. Aveiro, Portugal
A. Tomé, Univ. Aveiro, Portugal

Lecture Notes in Bioinformatics

Vol. 3745: J.L. Oliveira, V. Maojo, F. Martin-Sanchez, A. Sousa Pereira (Eds.), Biological and Medical Data Analysis. XII, 402 pages. 2005.

Vol. 3695: M.R. Berthold, R. Glen, K. Diederichs, O. Kohlbacher, I. Fischer (Eds.), Computational Life Sciences. XI, 277 pages. 2005.

Vol. 3692: R. Casadio, G. Myers (Eds.), Algorithms in Bioinformatics. X, 436 pages. 2005.

Vol. 3680: C. Priami, A. Zelikovsky (Eds.), Transactions on Computational Systems Biology II. IX, 153 pages. 2005.

Vol. 3678: A. McLysaght, D.H. Huson (Eds.), Comparative Genomics. VIII, 167 pages. 2005.

Vol. 3615: B. Ludäscher, L. Raschid (Eds.), Data Integration in the Life Sciences. XII, 344 pages. 2005.

Vol. 3594: J.C. Setubal, S. Verjovski-Almeida (Eds.), Advances in Bioinformatics and Computational Biology. XIV, 258 pages. 2005.

Vol. 3500: S. Miyano, J. Mesirov, S. Kasif, S. Istrail, P. Pevzner, M. Waterman (Eds.), Research in Computational Molecular Biology. XVII, 632 pages. 2005.

Vol. 3388: J. Lagergren (Ed.), Comparative Genomics. VII, 133 pages. 2005.

Vol. 3380: C. Priami (Ed.), Transactions on Computational Systems Biology I. IX, 111 pages. 2005.

Vol. 3370: A. Konagaya, K. Satou (Eds.), Grid Computing in Life Science. X, 188 pages. 2005.

Vol. 3318: E. Eskin, C. Workman (Eds.), Regulatory Genomics. VII, 115 pages. 2005.

Vol. 3240: I. Jonassen, J. Kim (Eds.), Algorithms in Bioinformatics. IX, 476 pages. 2004.

Vol. 3082: V. Danos, V. Schachter (Eds.), Computational Methods in Systems Biology. IX, 280 pages. 2005.

Vol. 2994: E. Rahm (Ed.), Data Integration in the Life Sciences. X, 221 pages. 2004.

Vol. 2983: S. Istrail, M.S. Waterman, A. Clark (Eds.), Computational Methods for SNPs and Haplotype Inference. IX, 153 pages. 2004.

Vol. 2812: G. Benson, R.D. M. Page (Eds.), Algorithms in Bioinformatics. X, 528 pages. 2003.

Vol. 2666: C. Guerra, S. Istrail (Eds.), Mathematical Methods for Protein Structure Analysis and Design. XI, 157 pages. 2003.

Table of Contents

Medical Databases and Information Systems

Data Analysis and Image Processing

Knowledge Discovery and Data Mining

Statistical Methods and Tools for Biomedical Data Analysis

Decision Support Systems

Collaborative Systems in Biomedical Informatics

Bioinformatics: Computational Models

Bioinformatics: Structural Analysis

Bioinformatics: Microarray Data Analysis

Application of Three-Level Handprinted Documents Recognition in Medical Information Systems

Jerzy Sas[1] and Marek Kurzynski[2]

[1] Wroclaw University of Technology, Institute of Applied Informatics,
Wyb. Wyspianskiego 27, 50-370 Wroclaw, Poland
jerzy.sas@pwr.wroc.pl
[2] Wroclaw University of Technology, Faculty of Electronics,
Chair of Systems and Computer Networks,
Wyb. Wyspianskiego 27, 50-370 Wroclaw, Poland
marek.kurzynski@pwr.wroc.pl

Abstract. In this paper the application of novel three-level recognition concept to processing of some structured documents (forms) in medical information systems is presented. The recognition process is decomposed into three levels: character recognition, word recognition and form contents recognition. On the word and form contents level the probabilistic lexicons are available. The decision on the word level is performed using results of character classification based on a character image analysis and probabilistic lexicon treated as a special kind of soft classifier. The novel approach to combining these both classifiers is proposed, where fusion procedure interleaves soft outcomes of both classifiers so as to obtain the best recognition quality. Similar approach is applied on the semantic level with combining soft outcomes of word classifier and probabilistic form lexicon. Proposed algorithms were experimentally applied in medical information system and results of automatic classification of laboratory test order forms obtained on the real data are described.

1 Introduction

Automatic analysis of handwritten forms is useful in such applications where direct information insertion into the computer system is not possible or inconvenient. Such situation appears frequently in hospital medical information systems, where physicians or medical staff not always can enter the information directly at the system terminal. Form scanning is considered to be especially useful in laboratory support software, where paper forms are still frequently used as a medium for laboratory test orders representation. Hence, in many commercially available medical laboratory systems a scanning and recognition module is available.

Typical form being considered here has precisely defined structure. It consists of separated data fields, which in turn consist of character fields. In our approach we assume that the whole form contents describes an object from the finite set of items and the ultimate aim of form recognition is selecting of relatively small

J.L. Oliveira et al. (Eds.): ISBMDA 2005, LNBI 3745, pp. 1–12, 2005.

subset of objects. Therefore, instead of using the classic pattern recognition approach consisting in indicating a single class, we will apply "soft" recognizer ([3]) which fetches the vector of soft labels of classes, i.e. values of classifying function.

In order to improve the overall form recognition quality, compound recognition methods are applied. Two most widely used categories of compound methods consist in combining classifiers based on different recognition algorithms and different feature sets ([4]). Another approach divides the recognition process into levels in such a way, that the results of classification on lower level are used as features on the upper level ([2]). Two-level approach is typical in handwriting recognition, in which the separate characters are recognized on the lower level and next on the upper level the words are recognized, usually with the use of lexicons.

In this paper, the method which uses both classifier combination and multilevel recognition is described. Probabilistic properties of lexicon and character classifier are typically used to build Hidden Markov Model(HMM) of the language ([11]). We propose another approach to the word recognition, in which probabilistic lexicon is treated as a special kind of classifier based on a word length, and next result of its activity is combined with soft outcomes of character classifier based on recognition of character image. Soft outcomes of a word classifier can be used next as data for semantic level classifier, which - similarly as previously - combined with object lexicon - recognizes the object described by the whole form.

The contents of the work are as follows. Section 2 introduces necessary background. In section 3 the classification methods on successive levels of object recognition problem are presented and concept of fusion strategies of character-based and lexicon-based classifiers are discussed. The proposed algorithms were practically implemented in application for automatic processing of laboratory test order forms in hospital information system. The system architecture and some implementation details are described in section 4. Results of experiments on proposed method efficiency are presented in section 5

2 Preliminaries

Let us consider a paper form F designed to be filled by handwritten characters. The form consists of data fields. Each data field contains a sequence of characters of limited length coming from the alphabet $\mathcal{A} = \{c_1, c_2, ..., c_L\}$. We assume that the actual length of filled part of data field can be faultlessly determined. The set $\mathcal{A}$ can be different for each field. Typically we deal with fields that can contain only digits, letters or both of them. For each data field there exists a probabilistic lexicon $\mathcal{L}$. Lexicon contains words that can appear in the data field and their probabilities:

$$\mathcal{L} = \{(W_1, p_1), (W_2, p_2), ..., (W_N, p_N)\}, \tag{1}$$

where W_j is the word consisting of characters from $\mathcal{A}$, p_j is its probability and N is the number of words in the lexicon.

The completely filled form describes an object (e.g. a patient in medical applications) and the data items written in the data fields are its attributes. The form contents, after manual verification is entered to the database, which also contains the information about the objects appearance probability. An example can be a medical information system database, where the forms contain test orders for patients registered in the database. The patients suffering from chronic diseases are more frequently examined, so it is more probable that the form being recognized concerns such a patient. Thus, this data base can be treated as a kind of probabilistic lexicon containing objects recognized in the past and the information about probability of its appearance, viz.

$$\mathcal{L}_{\mathcal{B}} = \{(b_1, \pi_1), (b_2, \pi_2), ..., (b_M, \pi_M)\}. \tag{2}$$

Our aim is to recognize the object $b \in \mathcal{L}_{\mathcal{B}}$ on the base of scanned image of a form F and both lexicons (1), (2). The recognition process can be divided into three levels, naturally corresponding to the three-level form structure:

- character (alphabetical) level – where separate characters are recognized,
- word level – where the contents of data fields is recognized, based on the alphabetical level classification results, their probabilistic properties and probabilistic lexicon (1),
- semantic level – where the relations between fields of the form being processed and lexicon (2) are taken into account to further improve the recognition performance.

In the next section the classification methods used on the successive levels of recognition procedure are described in details.

3 Three-Level Form Recognition Method

3.1 Character Recognition on the Alphabetical Level

We assume that on character (alphabetical) level classifier Ψ_C is given which gets character image x as its input and assigns it to a class (character label) c from $\mathcal{A}$, i.e., $\Psi_C(x) = c$. Alternatively, we may define the classifier output to be a L-dimensional vector with supports for the characters from $\mathcal{A}$ ([4]), i.e.

$$\Psi_C(x) = [d_1(x), d_2(x), ..., d_L(x)]^T. \tag{3}$$

Without loss of generality we can restrict $d_i(x)$ within the interval $[0, 1]$ and additionally $\sum_i d_i(x) = 1$. Thus, $d_i(x)$ is the degree of support given by classifier Ψ_C to the hypothesis that image x represents character $c_i \in \mathcal{A}$. If a crisp decision is needed we can use the maximum membership rule for soft outputs (3), viz.

$$\Psi_C(x) = arg\,(\max_i\ d_i(x)). \tag{4}$$

We have applied MLP-based classifier on this level. The vector of support values in (3) is the normalized output of MLP obtained by clipping network output values to $[0, 1]$ range and by normalizing their sum to 1.0.

Independently of nature of classifier Ψ_C, support vector (3) is usually interpreted as an estimate of *posterior* probabilities of classes (characters) provided that observation x is given ([4], [9], [10]), i.e. in next considerations we adopt:

$$d_i(x) = P(c_i \mid x), \quad c_i \in \mathcal{A}. \tag{5}$$

3.2 Data Field Recognition on the Word Level

Let the length $\mid W \mid$ of currently recognized word $W \in \mathcal{L}$ be equal to n. This fact defines the probabilistic sublexicon $\mathcal{L}_n$

$$\mathcal{L}_n = \{(W_k, q_k)_{k=1}^{N_n} : W_k \in \mathcal{L}, \mid W_k \mid = n\}, \tag{6}$$

i.e. the subset of $\mathcal{L}$ with modified probabilities of words:

$$q_k = P(W_k / \mid W_k \mid = n) = \frac{p_k}{\sum_{j:|W_j|=n} p_j}. \tag{7}$$

The sublexicon (6) can be also considered as a soft classifier Ψ_L which maps feature space $\{\mid W_k \mid : W_k \in \mathcal{L}\}$ into the product $[0,1]^{N_n}$, i.e. for each word length n produces the vector of supports to words from $\mathcal{L}_n$, namely

$$\Psi_L(n) = [q_1, q_2, ..., q_{N_n}]^T. \tag{8}$$

Let suppose next, that classifier Ψ_C, applied n times on the character level, on the base of character images $X_n = (x_1, x_2, ..., x_n)$, has produced the sequence of character supports (3) for the whole recognized word, which can be organized into the following matrix of supports, or matrix of *posterior* probabilities (5):

$$D_n(X_n) = \begin{pmatrix} d_{11}(x_1) & d_{12}(x_1) & \dots & d_{1L}(x_1) \\ d_{21}(x_2) & d_{22}(x_2) & \dots & d_{2L}(x_2) \\ \vdots & \vdots & \dots & \vdots \\ d_{n1}(x_n) & d_{n2}(x_n) & \dots & d_{nL}(x_n) \end{pmatrix}. \tag{9}$$

Now our purpose is to built soft classifier Ψ_W (let us call it *Combined Word Algorithm* - CWA) for word recognition as a fusion of activity of both lexicon-based Ψ_L and character-based classifier Ψ_C:

$$\Psi_W(\Psi_C, \Psi_L) = \Psi_W(D_n, \mathcal{L}_n) = [s_1, s_2, ..., s_{N_n}]^T, \tag{10}$$

which will produce support vector for all words from sublexicon $\mathcal{L}_n$.

Let $\mathcal{N} = \{1, 2, ..., n\}$ be the set of numbers of character positions in a word $W \in \mathcal{L}_n$ and $\mathcal{I}$ denotes a subset of $\mathcal{N}$. In the proposed fusion method with "interleaving" first the algorithm Ψ_C applied for recognition of characters on positions $\mathcal{I}$ on the base of set of images $X^{\mathcal{I}} = \{x_k : k \in \mathcal{I}\}$, produces matrix of supports $D^{\mathcal{I}}$ and next - using these results of classification - the lexicon $\mathcal{L}_n$ (or algorithm Ψ_L) is applied for recognition of a whole word W.

The main problem of proposed method consists in an appriopriate division of $\mathcal{N}$ into sets $\mathcal{I}$ and $\bar{\mathcal{I}}$ (complement of $\mathcal{I}$). Intuitively, subset $\mathcal{I}$ should contain these positions for which character recognition algorithm gives the most reliable results. In other words division of $\mathcal{N}$ should lead to the best result of classification accuracy of a whole word. Thus, subset $\mathcal{I}$ can be determined as a solution of an appropriate optimization problem.

Let $W^{\mathcal{I}} = \{c_{i_k} : k \in \mathcal{I},\ c_{i_k} \in \mathcal{A}\}$ be any set of characters on positions $\mathcal{I}$. Then we have following posterior probability:

$$P(W^{\mathcal{I}} \mid X^{\mathcal{I}}) = \prod_{k \in \mathcal{I}} d_{k\, i_k}(x_k). \tag{11}$$

The formula (11) gives conditional probability of hypothesis that on positions $\mathcal{I}$ of word to be recognized are characters $W^{\mathcal{I}}$ provided that set of character images $X^{\mathcal{I}}$ has been observed.

Applying for remaining part of the word sublexicon $\mathcal{L}_n$, we can calculate conditional probability of the whole word $W_j \in \mathcal{L}_n$, which constitutes the support (10) for word W_j of soft classifier Ψ_W:

$$s_j = P(W_j \mid X^{\mathcal{I}}) = P(W^{\mathcal{I}} \mid X^{\mathcal{I}})\, P(W_j \mid W^{\mathcal{I}}). \tag{12}$$

The first factor in (12) is given by (11) whereas the second one can be calculated as follows:

$$P(W_j \mid W^{\mathcal{I}}) = \frac{q_j}{\sum_{j:W_j\, contains\, W^{\mathcal{I}}} q_j}. \tag{13}$$

Since the support vector (12) of the rule Ψ_W strongly depends on the set $\mathcal{I}$ hence we can formulate the following optimization problem:

It is neccesary to find such a subset $\mathcal{I}^*$ of $\mathcal{N}$ and such a set of charcters $W^{\mathcal{I}^*}$ which maximize the maximum value of decision supports dependent on sets $\mathcal{I}$ and $W^{\mathcal{I}}$, namely

$$Q(\Psi_W^*) = \max_{\mathcal{I}, W^{\mathcal{I}}} \max_{j=1,2,\ldots,N_n} s_j(\mathcal{I}, W^{\mathcal{I}}). \tag{14}$$

The detailed description of suboptimal solution of the problem (14) which was applied in further experimental investigations can be find in [8].

3.3 Complete Form Recognition on the Semantic Level

For recognition of the whole form (object) on the semantic level we propose procedure called *Combined Semantic Algorithm* (CSA), which is fully analogous to the approach applied on the word level, i.e. relation between word classifier Ψ_W and probabilistic lexicon (2) is exactly the same as relation between the character recognizer Ψ_C and word lexicon (1). In other words, the form lexicon is treated as a special kind of classifier producing the vector of form supports (probabilities)

$$\pi = (\pi_1, \pi_2, \ldots, \pi_M), \tag{15}$$

which next are combined with soft outcomes (10) of word classifier Ψ_W.